Minor Revision

Thank you very much for your effort to further improve the manuscript.
The remaining issues seem to have been focused on the age analysis. The reviewer was, very thankfully, extremely prompt in providing comments on the manuscript. I regret that I have to ask you to revise the manuscript several times, but please revise it again so that the proper analysis can be conducted.

Yuki Yamada, Recommender

Dear Prof Yamada,

Thanks again for the comments of the reviewer and the editor. We have addressed Dr. Dienes's comments carefully, please see the detailed point-by-point response below and the changed traces in our manuscript.

We hope you will find the revised manuscript acceptable for in-principle acceptance.

Sincerely,

Hu Chuan-Peng,
Lei Yue

Reviews

Reviewed by Zoltan Dienes, 23 Mar 2023 11:56

The authors show they actually can pick up reasonable proportion differences using counts (which is what should be used for a multinomial). But for the age case, they have misunderstood me: I in no way meant they should use a t-test. I meant they should indicate what sort of effects thier procedure can pick up, using the same logic as they have e..g used here: "an article reported 30 participants, with age = 23.3 ± 3.5, we estimate the approximate number of participants under 20 is 5 (r code: `round((pnorm(20, mean = 23.3, sd =3.5) * 30))`), the number of participants aged between 21 ~ 30 is 24 (r code: `round((pnorm(30, mean = 23.3, sd =3.5) * 30)) - 5`), and participant aged between 30 ~ 40 will be 1." Sure a multinomial analysis can pick up more than just mean differences in age. But one still needs to show the model of H1 as a uniform has sensible properties, given such a model does not reflect a relevant theory. So one way of showing the sensitivity of the analysis is to imagine a diference only in means of a certain amount, translate that into the age bins, run the multinomial analysis, and see what it can pick up. At the moment all we know - from what is provided in the actual paper - is that it can get evidence for H1 when the sample data are generated from a uniform; but that is such an unlikely position for the real data to be in that it does not tell us much. Sorry to go on, but my point has not been addressed.

Response 1: Thanks a lot for the quick response and clarification. After a careful reading of your comments, we conducted the sensitivity analysis as you suggested (and sorry for the misunderstanding in our previous revision).

In this new simulation for age bins, we tested the minimal mean age difference that can be detected by the current setting of our Bayesian multinomial test (with non-informative prior and sample size = 1200, the justification for this sample can be found simulation in sex ratio). The simulation was conducted as follows:

Step 1: We estimated the mean (23.01) and *SD* (4.65) of Chinese participants' age from Jones et al (2021). These data will provide a starting point for simulation.

Step 2: We use the mean age 23.01 as the mean age of observed, and we choose a mean age difference (e.g., diff_age = 1 year), the mean age of the second group will be 23.01 + diff_age. We assume that both groups share the same *SD*.

Step 3: We generate multinomial data as mentioned by the reviewer, which is the way we will generate count data for different age bins when individual data is not available. For example, we generate the count number of participants under 18 by using `round((pnorm(18, mean = Mean1, sd =SD1) * 1200))`, and the count number of participants between 18 and 26 by using `round((pnorm(26, mean = Mean1, sd =SD1) * 1200)) - round((pnorm(18, mean = Mean1, sd =SD1) * 1200))`. Replace "Mean1" and "SD1" we get the data for group 1 and group 2. If the count data is zero for certain age bin, we replace zero with one to avoid error when calculating the Bayes factor.

Step 4: We conduct the Bayesian multinomial test, using data from group 1 as the observed and data from group 2 as the expected. We save the BF_{10} and $logBF_{10}$ resulting from this step.

Step 5: We repeated the Step 2 to Step 4, iterating the diff_age from 0 to 2, with the interval of 0.05. The results were as below:

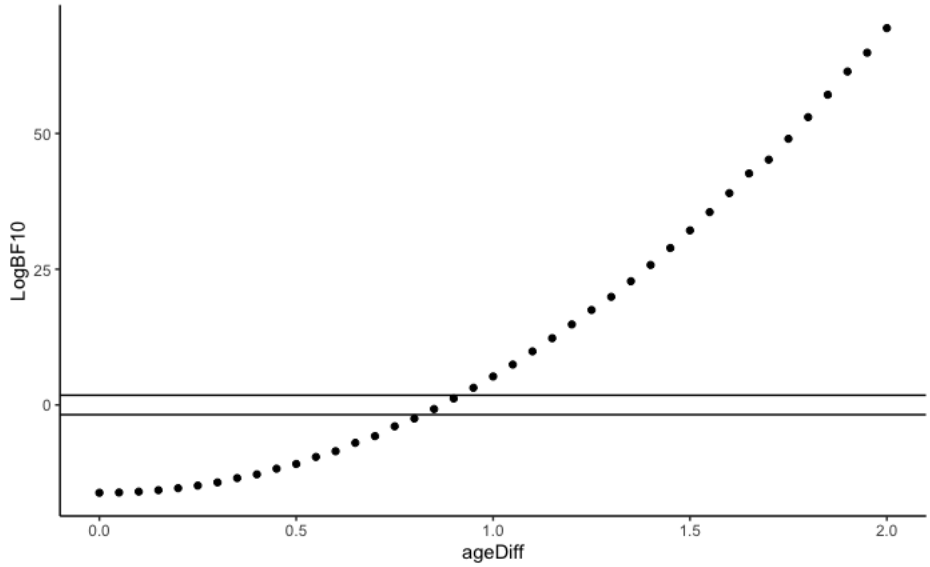

Figure R1. Log BF_{10} (y-axis) for different mean age differences (x-axis). The two black horizontal lines represent $logBF_{10} > log(6)$ (upper) and $logBF_{10} < log(1/6)$ (lower).

Please note that the mean age difference in our simulation, 0.05 to 2, corresponds to Cohen's *d* from 0.0107 to 0.429, if we assume the age is normally distributed. A mean age difference of 0.95 (Cohen's $d = 0.2041$) is the minimal effect that the current setting of the Bayesian multinomial test can detect with a threshold of BF >= 6 ($logBF_{10} > 2.3$).

We repeated the above Step 2 to Step 5 with different mean age of group 1 from 10 to 60, with 5 years as the interval, and fixed the *SD* as 4.65, we found that when the means age is moving toward two ends (0 or 60), the current setting will no longer be able to detect the mean age difference. However, this is not necessarily because the Bayesian multinomial test is not sensitive enough, instead, it is because most of the generated count data will be on both ends.

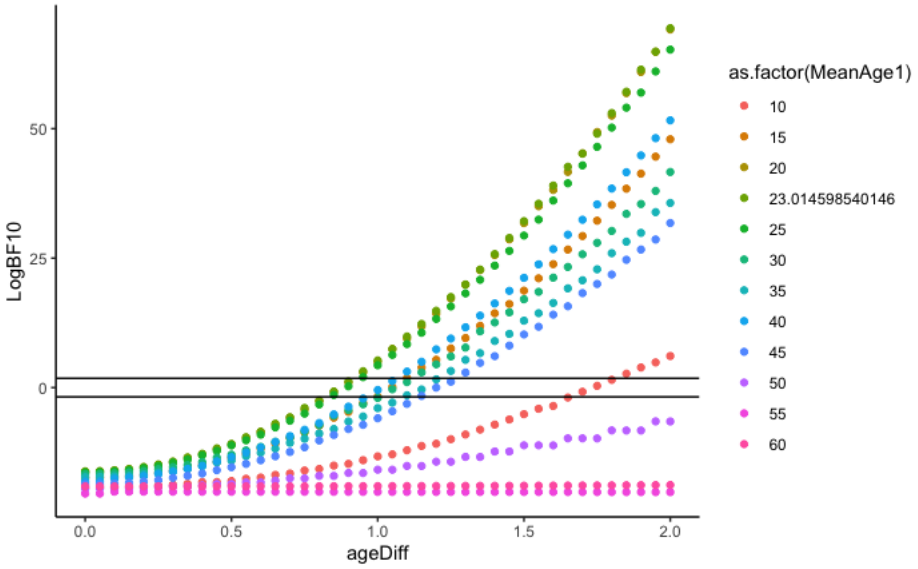

Figure R2.

In sum, the above simulation suggests that our Bayesian Multinomial test with non-informative prior can detect a mean age difference of 0.95 to 1.35, when the mean age is between 15 years old and 45 years old.

These mean ages, from 15 to 45, represented the majority of studies. Our revisit to Rad et al (2018) found that 21 out of 33 studies investigated young and mid-aged adults, with mean age range from 18.63 to 36.81.

Note that the simulation above lacks of stochastic process, each mean age difference only returns one BF value. To further confirm the above result, we used an alternative way to conduct the simulation. In this approach, we generate age data and convert these age data into age bins, instead of generating count data for each age bin directly.

That is, we replace Step 3 by first generating 1200 age data by using truncated normal distribution `rtruncnorm(n=sampe_N, a=0, b =100, mean = Mean1, sd = SD1)`, and then we count the number of ages data that fall in each age bin. After that, we conduct the Bayesian multinomial test.

We also used the different mean ages and different mean differences between age groups. As in Figure R3, this simulation gets similar results as the above simulation approach, with a bit more noise. From the data, found that, when the mean age is from 15 to 45, the current setting can detect a mean age difference from 1.1 years to 1.6 years, corresponding to Cohen's d of 0.236 to 0.344.

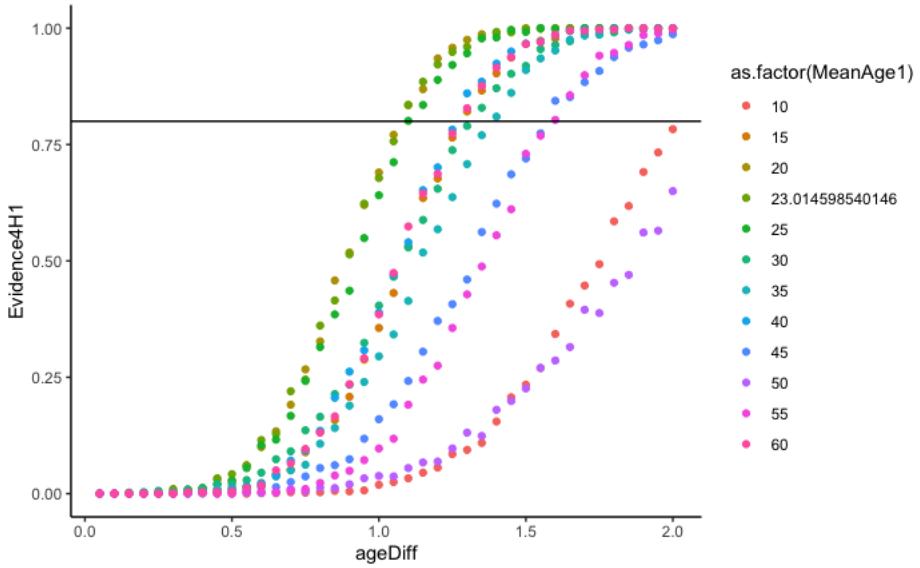

Figure R3. Simulation results with sample size N = 1200. The x-axis the differences between the means of group 2's age and group 1's age, different colors represent different means ages from group 1. The y-axis is the percentage of BF >= 6 in 1000 simulations.

In conclusion, our sensitivity analysis revealed that with a sample size of 1200 and non-informative prior, the current Bayesian multinomial test can detect a mean difference of 0.95 years (in the best case) to 1.6 years (in the worst case) when participants are young or mid-aged adults. As the final sample size will be larger than 1200, the current setting will be sensitive to detect the even smaller difference.

Again, we appreciate your efforts in reviewing our RR and hope the simulations above addressed your concern about the sensitivity of our statistical test for comparing age distributions.